Article

Not peer-reviewed version

Advanced Identification of Prosodic Boundaries, Speakers and Accents Through Multi-Task Audio Pre-Processing and Speech Language Models

Fco. Javier Lima Florido and Gloria Corpas Pastor *

Posted Date: 10 February 2025

doi: 10.20944/preprints202502.0579.v1

Keywords: speech processing; prosodic boundaries detection; speaker change detection; accent classification; Transformer architecture; Wav2Vec2; NLP; multi-task classification

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Advanced Identification of Prosodic Boundaries, Speakers and Accents Through Multi-Task Audio Pre-Processing and Speech Language Models

Francisco Javier Lima Florido and Gloria Corpas Pastor *

IUITLM, University of Malaga; fco.javier.lima@uma.es

* Correspondence: gcorpas@uma.es

Abstract: In recent years the advances in deep neural networks (DNNs) and large language models (LLMs) have led to major breakthroughs and new levels of performance in Natural Language Processing (NLP), including speech processing related tasks. Based on these new trends, new models such as Whisper and Wav2Vec 2.0 achieve robust performance in speech processing tasks, even in speech-to-text translation and end-to-end speech translation, far exceeding all previous results. Although these models have shown excellent results in real-time speech processing, they still have some accuracy issues for some tasks and high latency problems when working with large amounts of audio data. In addition, many of them need audio to be segmented and labelled for speech synthesis and annotation tasks. Speaker diarisation, background noise detection, prosodic boundary detection and accent classification are some of the pre-processing tasks required in these cases. In this study we will fine-tune a small Wav2Vec 2.0 base model for multi-task classification and audio segmentation. A corpus of spoken American English will be used for the experiments. We intend to explore this new approach and, more specifically, the performance of the model with regards to prosodic boundaries detection for audio segmentation, and advanced accent identification.

Keywords: speech processing; prosodic boundaries detection; speaker change detection; accent classification; Transformer architecture; Wav2Vec2; NLP; multi-task classification

1. Introduction

Speech is the basic form of communication that humans use every day. Because of this, the possibility to communicate with machines through speech has been an interest for many decades. Over the past few years, language and speech models have acquired significant importance in society, mainly due to the recent advances in Natural Language Processing (NLP) and deep neural networks (DNNs). Current models, like Whisper [1] and Wav2Vec 2.0 [2] allow users to have fluent communication with conversational bots, either over text or speech, in real time. The improvements in speech processing have also been reflected in many other system features, for instance, automatic captioning of speech generation for artificial text readers.

Despite these improvements, there are still many scenarios where speech processing needs improvement. Automatic Speech Recognition (ASR) has achieved high scores in performance for common languages, but for low-resource languages, the performance of ASR models is still far from acceptable. In the case of the Arabic language, the word error rate is around 25% or even higher in some cases [3,4]. As demonstrated in [5],to improve their performance, ASR models can be trained applying pre-processing techniques such as segmentation, word modeling or detection of different environmental or speaker variations. These tasks are modeled as speech classification tasks and can be easy to address when sufficient labelled data is available. However, for many languages, it is hard to find enough transcribed and labelled data to train a robust ASR model [6].

In this work, we propose a pre-processing speech model that can perform three tasks at once: prosodic boundary detection (PBD), speaker change detection (SCD) and accent classification (AC).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

Our hypothesis is two-fold: (i) that a Wav2Vec 2.0 base model, as a feature extractor block, can be fine-tuned to extract the features from the audio that are needed to perform the three classifications tasks correctly, and that (ii) this transformer block can generate context representations that share all these features. Thus, the most distinctive feature of our model is, precisely, that it is trained only on audio data, which can help to improve speech recognition performance in terms of precision and resource optimisation.

The rest of the text is organised as follows. Section 2 presents a brief overview about past and recent methods for speech processing. Section 3 describes the methodology and materials used in our experiments. Section 4 reports the experimental results, which are then discussed in Section 5. Finally, Section 6 provides the main conclusions and directions for future work.

2. A Brief Overview

In computer science, speech processing has been defined as the study of speech signals and their processing methods. It encompasses a range of tasks, including classification, recognition, transformation, and generation of speech data, among others. Speech processing tasks have traditionally been based on signal processing [7,8] and linguistic features [9–11] until the growth of Deep Learning (DL) [12–15] In this section, we focus on DL-based techniques applied to three tasks which are particularly relevant to our research: phrase boundary detection (PBD), speaker change detection (SCD) and accent classification (AC). As a convenient theoretical framework, a short description of the techniques applied to these tasks before and after the popularity of DNNs will be presented in order to locate this study in context.

Before Deep Learning, Hidden Markov Models (HMM) [8,16] were currently applied for speech recognition (ASR) and became the dominant paradigm until DNNs started to outperform them [17,18]. Other speech tasks, such as voice activity detection (VAD), speaker recognition and diarisation, prosodic detection and segmentation, among others, play an important role as speech pre-processing tasks to ensure optimal performance in ASR and speech generation (text-to-speech, TTS). Regarding these tasks, numerous alternative methods have been employed prior to the emergence of DNNs. Signal processing algorithms and linguistic feature methods have demonstrated a notable performance for VAD [7], speaker diarisation [19], classification problems [20], and prosodic-related tasks [10].

Even though some of the former methods and algorithms are still in use [21,22], DNNs are currently the most popular methods for speech pre-processing, as well as for ASR, to the point where most of the new methods for these tasks are based on or derived from DNNs (see [23] for a comprehensive review).

2.1 Speech Processing Methods and Techniques

As DNNs became the standard for ASR tasks, the encoder-decoder architecture obtained the most promising performance. In this approach, the encoder extracts features from the input speech, while the decoder transforms those features into the desired output, namely a transcription [24]. In this manner, the encoder learns to generate representations or embeddings that represent the extracted features. A number of different methods of representation is employed in the literature for speech processing tasks. For instance, i-vectors [25], d-vectors [26] and x-vectors [27] are convolutional NNs (CNN) and recurrent NNs (RNN)-based embeddings that have resulted in improved performance of tasks such as speaker diarisation or VAD. However, Transformer NNs [28] have surpassed RNNs and CNNs in the context of speech processing tasks. The utilisation of self- and cross-attention Transformer NNs allows parallelisation and performance improvements that explain the success of them in Natural Language Processing (NLP) tasks [29].

2.2 Automatic Speech Segmentation

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

Despite the mentioned computational progress, automatic speech segmentation is still one of the most difficult challenges in speech processing. The variation in speaker characteristics, speaking style and environment make the learning process of prosodic boundaries hard for machines [30]. However, detecting and processing prosodic phrasing is relevant for many speech processing tasks as it can help to improve naturalness in speech synthesis and to enhance audio data for training [31]. In the case of speech synthesis, the common scenario for the detection of prosodic boundaries is performed by relying solely on text [32], while other systems rely solely on acoustic information [33], or on a combination of both lexical and acoustic information [21]. For the purpose of this study, we have opted for the detection of prosodic boundaries based solely on acoustic information (cf. section 3. Materials and methods).

Speaker change detection (SCD) is another speech segmentation task used to determine the boundaries between speakers in a conversation. In simpler words, SCD is the action to detect the moments when a speaker stops talking and another speaker starts to talk, regardless of the identity of the speakers. SCD is part of speaker diarisation systems (who talks and when) [34] but it is also useful for speaker tracking in video data [35] and for transcribing audio with multiple speakers [36]. There are many approaches to perform this task based on DNNs such as the previously mentioned x-vectors [37] or methods based on self-attention mechanisms [38,39]. Overlapped segments (i.e. several speakers talking at the same time) are a common, recurrent problem in SCD. Many of the existing methods opt for the omission of the overlapped segments. However, similar DNNs-based approaches are being applied specifically to overlapped speech detection [40–42]. Besides, adding extra information during training through multitasking techniques has shown that multitask models can reach better results than models trained for only one task [43].

2.4 Accent Identification

Before DNNs, studies in accent identification focused on leveraging linguistic features in statistical approaches. For instance, some of the features that have been combined with statistical analysis are prosodic parameters, syllable structure and speaker characteristics [20,44,45]. Recent studies are centered on DNNs for spoken language identification [46], the generation of representations around features related to accent identification [47–49] or other DL approaches for accent classification [50–52]. Other recent approaches have explored the possibilities of combining accent identification and ASR as a viable way to improve the performance of ASR models [53]. It is notable to mention that some methods, like the one presented by Ghorbani and Hansen [54], have been able to reach better scores than human-perceived scores.

2.5 Joint and Multitask Learning in Speech Processing

Joint learning and multitask learning are both approaches used in machine learning to leverage shared information across tasks. In speech processing, this approach has demonstrated to be beneficial and to get better performance in many tasks [55]. The method presented in [56] shows that joint training could help the encoder of the NN to learn transferable, robust, and problem-agnostic features that carry on relevant information from the speech signal, which contributes to discovering general representations. For speech segmentation, joint and multitask frameworks have been applied to prosodic boundaries detection and SCD with successful results [33,43,57]. With regards to joint learning methods which involve accent classification, recent studies have shown high accent classification performance by adding this task to speech recognition models [58–61].

3. Material and Methods

The main goal of the research presented in this paper is to develop an innovative multitask training approach for PBD, SCD and accent classification by relying only on acoustic information. To address this multitask problem, we have fine-tuned a Wav2Vec 2.0 model for audio frame classification with three classification layers on top of the model, one layer for each task. For this

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

experiment, we have used an American English oral corpus as a base dataset for training and validation of the models. This corpus has been pre-processed using different methods for each task, which has led to the generation of new datasets adapted for the purpose of the experiment. The following sections provide detailed information on the base dataset selected and the transformations applied to generate the final datasets, as well as the methods applied for the training, evaluation and validation of the models.

3.1 Speech Corpus Used in the Experiment

The selected corpus for the experiment is the Santa Barbara Corpus of Spoken American English (SBCSAE)[1] [62]. This is a prosodic annotated speech corpus (also named spoken corpus), which encompasses a total of 60 transcribed and annotated conversations, including the timestamps at the level of individual intonation units. The corpus is based on a large body of recordings of naturally occurring spoken interaction from all over the United States, representing a wide variety of people of different regional origins, ages, occupations, genders, and ethnic and social backgrounds. The predominant form of language use represented is face-to-face conversation, but the corpus also documents many other ways people use language in their everyday lives: telephone conversations, card games, food preparation, on-the-job talk, classroom lectures, sermons, story telling, town hall meetings, tour-guide spiels, and more. The transcriptions are formatted in one utterance per line, together with the initial and final timestamps, the name of the speaker and the intonation information of the utterance. In addition, the metadata of each conversation includes a short description and the location where it was recorded. All this information makes the SBCSAE a useful dataset for many supervised learning tasks in speech processing. Given the purpose of this work, we are specifically interested in the timestamps for SCD and PB, and the location information for the accent detection.

3.2 Data Preparation

This section covers all the automatic pre-processing and annotation tasks applied to the dataset used for the experiments. The two first subsections explain the design of the annotation process for each task, while the last one describes how all the processes have been applied automatically.

3.2.1 Audio Frame Labels

Audio segmentation can be seen as the act of slicing the audio into pieces following a specific criterion. Our approach aims to detect the specific frames where the speaker finishes an utterance or ends their intervention. For this reason, we needed to define the frames and assign specific labels for each frame. We have followed the methodology presented in [63] where they have trained a prosodic boundary detection (PBD) model for the Czech language. In their work, the authors assigned a label every 20ms of audio, giving a 1, if the frame contains a boundary, or a 0 if not. In addition, to avoid delays from the manual annotation, the frames around each prosodic boundary (PB) were given a label between 0 or 1, thus creating a linearly increasing and decreasing interval. For the PBD task, we have followed this exact same method applied to the SBCSAE.

The SBCSAE includes the name of the speaker for each utterance in the transcript file. We used this information to create the reference data for the SCD task in a similar manner to previously mentioned PB references: we have given the label 1 to the frames when a new speaker starts to talk and 0 for the next utterances of the same speaker. This method for label setting has one problem: it does not consider whether two or more speakers are talking at the same time. To address this issue, the SBCSAE corpus includes the labels “BOTH” and “MANY” when more than one person is speaking. This annotation has been adapted to our labelling method by assigning label 2 to the frames that correspond to these utterances. This is also known as Overlapping Speech Detection (OSD). In addition, the corpus also includes specific labels for timestamps where a sound or environmental

[1] The SBCSAE corpus can be accessed and downloaded here: https://www.linguistics.ucsb.edu/research/santa-barbara-corpus

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

noise occurs. Following our labelling process, we have assigned label 3 to these timestamps. In this way, label 3 indicates frames without voice activity. Although our dataset is focused on SCD, these annotations will enable the model to perform OSD and VAD also.

3.2.2 Accent Classification Annotation

The labelling process for accent detection differs considerably from the process applied to SCD and PBD (see above). The SBCSAE includes the city and state from the United States corresponding to each one of the conversations as part of the metadata of the audio. For this reason, it has been necessary to label the whole audio with its accent rather than splitting the audio into frames. This poses another challenge, that is, to merge the audio classification task with two audio frame classification tasks. To solve this, we have assigned the corresponding accent label to every frame of the audio and then reduced the set of frame labels to only one for the whole audio.

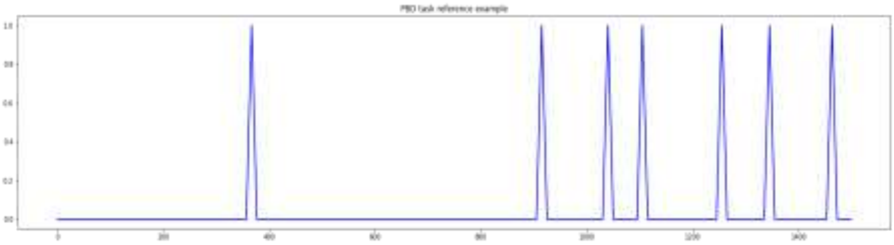
To establish the number of classes for this task, instead of assigning the labels based on the states, we have applied the classification of the major regional dialects of American English demarcated by Labov et al. [64]. This classification divides the map of the United States in six dialect regions. Table 1 shows the six dialects, the states that belong to each region and the labels that we have assigned for each dialect to prepare the data for this task.

Table 1. States included in the SBCSAE per dialect region, plus the label assigned to each region.

Regional dialects	**States**	**Label**
Inland North	Illinois, Wisconsin, Michigan	0
New England	Massachusetts, Vermont	1
Mid-Atlantic	Delaware, Pennsylvania,	2
South	Alabama, Texas, Kentucky, Louisiana	3
Midland	Indiana, Kansas	4
West	Montana, California, New Mexico, Arizona, Oregon, Washington, Nevada, Idaho, Colorado	5

3.2.3 Automatic Annotation Process

Once the labelling process is completed, the annotation files created from the SBCSAE are converted to the desired format (see sections 3.2.1. Audio frame labels and 3.2.2. Accent classification annotation). In order to automate the process, various Python scripts have been designed to streamline a variety of sequential tasks. The first task consists in splitting the audios into chunks of 30 seconds with a 15 second overlap between one audio and the next, to avoid context missing. Then, reference labels for PBD and SCD tasks are generated from the timestamps in the original annotations, and the speaker labels. The result is a set of 1500 labels for each task. Figure 1 shows example representations of the set of labels for PBD and SCD. The last script assigns the labels corresponding to the dialects according to the name of the state which appears in the metadata. The part of the scripts that perform the extraction of the key information were coded using regular expressions. The extracted data was saved as JSON files.



Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

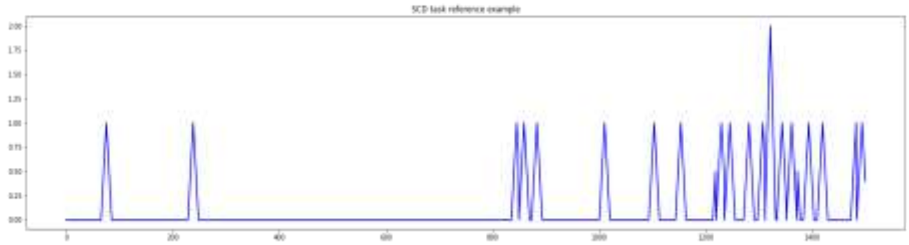


Figure 1. Examples of the reference label sets for PBD and SCD.

Once all the conversion and annotation processes described have been performed, the target dataset has been created through the *datasets* package from Huggingface[2]. This package allows the organisation of the dataset in three splits: train, test and validation. The entries inside each one of the splits contains the attributes "path" and "label". The "path" is the absolute path to the audio file and the "label" includes the three set of reference labels, one set for each task. After the generation of the dataset format, the entries are randomly divided into the three splits as follows: 80% of the dataset for training, 10% for testing and another 10% for validation.

3.3 Model Architecture and Experimental Design

Wav2Vec 2.0 (hereafter referred to as "wav2vec2") is a pretrained speech model whose architecture is mainly based on Transformers [2]. The model has a structure of three main parts: the feature encoder, the context network and the quantisation module. The encoder consists of several blocks of temporal convolutions that normalise the speech input and outline the number of time-steps of the speech which, in their turn, serve as input to the context network. The context network is a Transformer NN which learns contextualised representations from the output of the encoder. The quantisation module discretises the output of the feature encoder via product quantisation for self-supervised training. Figure 2 shows an illustration of the wav2vec2 framework. This model is able to learn representations from speech audio in the same way language models learn linguistic representations from text, achieving better results in ASR than previous state-of-the-art models [2].

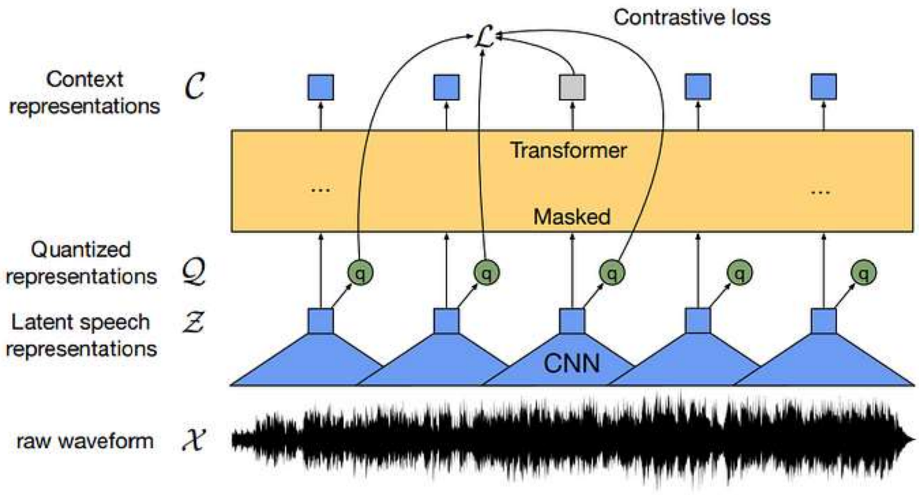


Figure 2: Illustration of the wav2vec2 framework by Baevski et al. *[2]*

In this study we have leveraged the language representations generated by wav2vec2 to perform a supervised learning experiment. This experiment has consisted in fine-tuning and validating a wav2vec2 model for multitask audio frame classification. To implement the desired architecture of

[2] https://huggingface.co/docs/datasets/index

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

the model, we have used the *transformers*[3] package by Huggingface, which already includes an implementation of an audio frame classification architecture using a wav2vec2 encoder and a classification head on top. Given that we need three classification heads instead of only one, we have built our own architecture for audio frame classification based on the existing implementation in the transformers package. Each one of the classification layers performs its own task: one of them for PBD, another for SCD and the last one for accent classification. During training, the context representations generated by the wav2vec2 block are given to each one of the classification heads and they produce their own results. This process is illustrated in Figure 3.

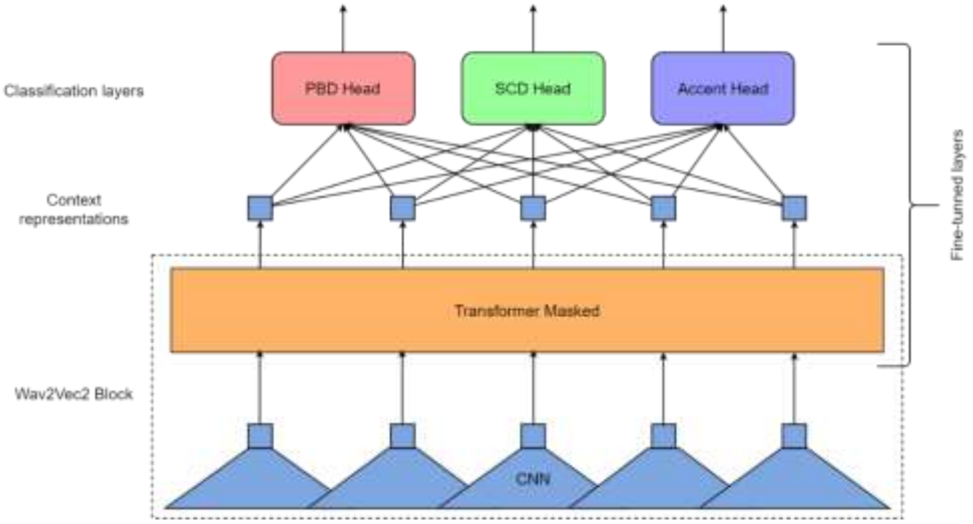


Figure 3. Model architecture and training process

3.4 Training, test and validation processes

Following previous studies (cf. section 2.5), we have focused on multitask training approaches as a key part of the training methodology. To train our model, we have used the wav2vec2 base model which can be obtained from the Huggingface models repository[4]. This is the wav2vec2 block of our architecture. As shown in Figure 3, only the Transformer layer of this block is fine-tuned during training. On the other hand, the three classification layers are randomly initiated and fully trained. The training is a supervised learning process, as the dataset includes the audio as inputs and the reference labels to adjust the loss function. For each step of the training, the context representations generated by the wav2vec2 block are given to the classification layers as the hidden states, and each one of the classification layers produces their corresponding label outputs. Following the starting hypothesis (cf. section 1. Introduction), the Transformer layer of the wav2vec2 block has been fine-tuned to extract the necessary features to perform the three classification tasks and generate the context embeddings that represent those features.

To achieve our goal, we have designed a composed loss function which is the combination of the loss from one of each classification layers. For PBD and SCD we have applied Mean-Squared Error (MSE) as the loss function, and Cross Entropy (CE) has been the loss function applied for Accent classification (AC). The reason for this difference in loss calculation is that for AC the labels are the same for every frame, therefore the output of the layer is converted into a probability distribution, instead of a set of labels. The composed loss is calculated as the weighted average of the three individual functions as follows:

$$\text{loss}=\text{loss}_{pbd}\,w_{pbd}+\text{loss}_{scd}\,w_{scd}+\text{loss}_{ad}\,w_{ad}. \quad (1)$$

[3] https://huggingface.co/docs/transformers/index

[4] https://huggingface.co/facebook/wav2vec2-base

The challenge of this loss function was to determine the optimal weights (w_{pbd}, w_{scd}, w_{ad}) applied to every task. To address this challenge, we have run a set of training experiments, using different parameters and weights. This process has helped to get the best context representations by fine-tuning the Transformer layer. The post-training results are presented below (cf. section 4. Results).

Next, the fine-tuned models have been evaluated through test and validation; for this task, the corresponding subsets from the dataset have been used. Testing has been done during fine-tuning, after each epoch. To this end, a composed metric has been used, similarly to what we did in the case of the loss function. However, we have used accuracy as a metric for accent classification instead of CE, and the MSE values from PBD and SCD have been inverted (1 - MSE). After training, the best checkpoint is selected based on the results of the test phases.

4. Results

Once the model has been fine-tuned, the validation is performed for each task separately by applying post-processing and different metrics. Post-processing consists in the reverse conversion of the label set of SCD and PBD classifiers by turning the labels into timestamps. After the reverse conversion, the F1 metric is applied to measure the performance of SCD and PBD tasks. In the case of the Accent classifier, its performance is measured using the accuracy metric (ACC).

4.1 Post-Processing Approach and Validation Measures

To convert the output labels given by the model back to timestamps, we applied different criteria to the PBD outputs and the SCD outputs. For the first task, the frames with label 1 are the PBs detected, although the model is not entirely accurate when assigning the labels. For that reason, we have established a threshold of 0.5. Every label bigger than this threshold is considered as a PB. In the case of SCD, the approach is similar, every frame with a label above 0.5 is considered a SC timestamp. We have also established thresholds for the labels corresponding to more than one person talking (1.5) and the labels for environmental sound or noises (2.5). For AC, no post-processing was needed.

After the post-processing was applied, we obtained two lists of detected timestamps for PBD and SCD respectively, and the corresponding references timestamps. To evaluate the performance of each task, we have applied several metrics, namely, Precision, Recall and F1. In the case of the AC output, we have measured the performance using ACC by comparing the reference set of accent labels and the output given by the Accent classifier.

4.2 Parameters Evaluation and Model Validation

The results discussed in the section are intended to fulfil a two-fold purpose. One is to validate the methodology and the trained models. The other is to find the optimal weights for the loss function. Tables 2-4 show the key results of each training, and their corresponding parameters, in terms of the applied loss weights and the number of epochs. For PBD and SCD task results, Table 2 (PBD) and Table 3 (SCD) contain the Precision, Recall and F1 values. In addition, Table 3 contains the Accuracy results for the Accent classification task.

Table 2. Validation results (%) for the PBD task

Models	Applied weights (w_{pbd}, w_{scd} and w_{ad})	Epochs	Precision	Recall	F1
1	0.38, 0.6, 0.02	10	63.49	50.43	56.21
2	0.33, 0.33, 0.33	20	67.19	44.24	53.35
3	0.4, 0.4, 0.2	20	66.66	60.67	57.58

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

4	0.4, 0.5, 0.1	20	68.83	49.82	57.8
5	0.38, 0.6, 0.02	20	69.56	58.3	64.44
6	0.39, 0.6, 0.01	50	75.29	**75.9**	75.59
7	0.399, 0.6, 0.001	50	**78.73**	75.01	**76.78**
8	0.299, 0.7, 0.001	50	76.02	71.21	73.54

Table 3. Validation results (%) for the SCD task

Model ID	Applied weights (w_{pbd}, w_{scd} and w_{ad})	Epochs	Precision	Recall	F1
1	0.38, 0.6, 0.02	10	63.2	42.8	51.04
2	0.33, 0.33, 0.33	20	61.43	33.6	43.45
3	0.4, 0.4, 0.2	20	62.43	38.16	47.37
4	0.4, 0.5, 0.1	20	63.07	43.88	51.76
5	0.38, 0.6, 0.02	20	62.18	57.49	59.74
6	0.39, 0.6, 0.01	50	70.89	72.72	71.79
7	0.399, 0.6, 0.001	50	69.74	**75.39**	72.45
8	0.299, 0.7, 0.001	50	**74.78**	74.46	**74.12**

Table 4. Validation results (%) for the Accent classification task

Model ID	Applied weights (w_{pbd}, w_{scd} and w_{ad})	Epochs	Accuracy
1	0.38, 0.6, 0.02	10	99.46
2	0.33, 0.33, 0.33	20	99.46
3	0.4, 0.4, 0.2	20	99.28
4	0.4, 0.5, 0.1	20	99.28
5	0.38, 0.6, 0.02	20	**99.82**
6	0.39, 0.6, 0.01	50	**99.82**
7	0.399, 0.6, 0.001	50	83.27
8	0.299, 0.7, 0.001	50	92.63

In the three tables, each entry corresponds to the results from the model trained using the indicated weights for the loss function and the number of epochs. Besides, entries are ordered by the number of epochs, since we have observed that this parameter has the higher impact on the results. This can be noticed in the results from models 1 and 5. Model 5 has achieved a higher punctuation in all three tasks, even though both models have been trained with the same weights. No improvements can be observed in learning for more than 50 epochs, which justifies the maximum number of epochs (50) presented in the tables.

The results obtained during the fine-tuning of the applied weights indicate that the task that presents most problems for the models to predict is SCD. This is not unexpected, as the references from SCD task present higher variations at labels per frame than the other two (see 3.4). On the contrary, the high accuracy levels achieved at the Accent classification task show that this is the task that presents less difficulties. Since the model only needs to predict one label for the AC task, the high difference in the results between this task and the other two is not surprising. This explains why the weight for Accent classification showed the highest decrement, whereas the weight for SCD is increased in each training by the highest factor (see 3.4).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

Finally, our findings suggest that model 8 is the one that exhibits the best performance. Even if the results for Accent classification and PBD are affected negatively by the weights of this model, it shows an appropriate balance in performance for the three tasks.

5. Discussion

From the perspective of previous studies, this is not the first time that wav2vec2 is leveraged to perform PBD in a single task environment or leveraged to perform multitask speech processing. In [63], the authors present a methodology to fine-tune a wav2vec2 model to detect PB on Czech speech data, relying solely on acoustic information. An extended approach based on the previous one, was also applied for multitask speech processing on English data in [57], where the trained model performs SCD, OSD and VAD as different tasks. Regarding other NNs architectures, a few studies have introduced multitask or joint learning approaches for SCD, PBD and AC [33,43,58]. In [30] fine-tuned Whisper models for PBD on the SBCSAE are presented with satisfactory results, although their acoustic model performance was worse than their lexical model. Referring to Accent classification, the work in [65] presents a particular Multi Kernel Extreme Learning Machine (MKELM) architecture which has achieved similar scores than other DNNs architectures.

Given the possibilities of Transformer models for speech processing, the objective of this work was to fine-tune a wav2vec2 model with three classification layers to perform PBD, SCD and AC by relying only on acoustic information. We have tested our method on spoken English by using the SBCSAE corpus, like most of the models analysed in this study. To the best of our knowledge, this is the first experiment that applies one single model to perform the three target tasks at the same time. The results obtained during the fine-tuning experiments show that our model can achieve a performance score close to previous methods or even higher.

A comparison between our model and the most relevant previously proposed models is highlighted in Table 5. This comparison shows that our model has achieved slightly lower performance scores for SCD and PBD than previous methods, but its performance on AC is the best.

Table 5. Summary of results from previous studies and our results

Models	PBD Score (% F1)	SCD Score (% F1)	AC Score (% Accuracy)
[63]	82.73	-	-
[57]	-	90.79	-
[33]	91	-	-
[43]	-	88.32	-
[58]	-	-	75.2
[30]	73	-	-
[65]	-	-	84.72
Our model	**73.54**	**74.12**	**92.63**

Our model exhibits lower performance for PBD than the models presented in [33] and [63], but higher than the model presented in [30]. In [63] the described method is applied to the Czech language and the model presented in [33] is trained on text information, not only on acoustics/audio data. This explains the difference in performance compared to our model. On the other hand, [30] is the only approach that has used the same corpus (SBCSAE). The performance of this model, which relies solely on acoustic information, is slightly lower than our model's performance. Regarding the SCD task, our model has the lowest performance score on the table, although the three models were trained on English corpora following multitask approaches. However, the architecture and the tasks of our model are more complex than the previous models, which can explain the lower results. Finally, for AC, our model shows better performance than previous works. The main differences in

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

this case are that the model in [65] is not a Transformer model and the approach applied in [58] combine ASR and AC.

In addition to the mentioned differences between our work and previous studies, it is important to note that, except the model described in [57], this is the only model that performs three speech processing tasks. In terms of model size, adding a new classification layer does not have an important impact, keeping the model size almost the same. This means that our model requires one-third of the memory compared to single-task models that perform the same three tasks, allowing for more efficient and cost-effective use of resources.

6. Conclusions and Future Work

Following our initial hypothesis, we assume that speech representations from the wav2vec2 framework could be leveraged in multitask speech processing. In this line, the main aim of this work was to demonstrate that a single wav2vec2 acoustic model can be fine-tuned, using joint learning techniques, for PBD, SCD and Accent classification. Our findings confirm our hypothesis: our model (trained and validated on the SBCSAE) is the best performing one, as it has achieved an F1 score of 0.74 for PBD and SCD, and a 0.93 Accuracy score for Accent classification.

To the best of our knowledge, ours is the first fine-tuned speech model which can perform all three PBD, SCD and Accent classification tasks by relying solely on acoustic information. This can be particularly useful in situations where no transcriptions or other language data are available. These tasks can be applied to speech pre-processing and segmentation, which can help to improve speech recognition performance in terms of precision and resource optimisation. This is the main contribution of this study. Another contribution is the data preparation process applied to the SBCSAE, which has been fully automated. This makes the process reproducible, scalable and transferable to other timestamp annotated corpora (with some minor adjustments, if needed). Finally, the proposed model architecture further advances the state-of-the-art fine-tuning methods, as the classifier layers can be trained to perform any audio frame classification task, which means that our model can be easily adapted to fine-tune models on other corpora.

The scope of this study was limited in terms of metrics and the relatively small data sample. It would be worth considering further complementary metrics, such as coverage, and measure training, as well as inference times of the models. But the most important limitation is related to the data used for training and validation. Our findings suggest that better results could be achieved by expanding the train and validation dataset with more speech corpora. Despite the significant contributions of our study, the issue of data size is an intriguing one which could be usefully explored in further research. To this end, we intend to repeat the experiments using more English datasets, in addition to the SBCSAE, to train and validate our model. In addition, several issues related to the limitations of our study remain to be solved. For instance, we will apply more metrics to expand the evaluation of the methods analysed. Finally, a natural progression for this work would be to test our model on other languages and on multilingual data (e.g., parallel corpora). In addition, our methodology could be applied to test other speech models and compare the results. Another possible study could be to leverage our model on cascaded or multi model speech systems, for instance, in a complete speech recognition system. This would be a fruitful area for further work.

Author Contributions: Conceptualisation, F.J.L. and G.C.; methodology, F.J.L. and G.C.; software, F.J.L.; validation, F.J.L.; formal analysis, F.J.L.; investigation, F.J.L. and G.C.; resources, G.C.; data curation, F.J.L.; writing—original draft preparation, F.J.L. and G.C.; writing—review and editing, G.C.; visualization, F.J.L.; supervision, G.C.; project administration, G.C.; funding acquisition, G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was carried out in the framework of various research projects on multilingual language technologies (VIP II, ref. no. **PID2020-112818GB-I00/AEI/10.13039/501100011033**), RECOVER, ref. no. **ProyExcel_00540;** DIFARMA, ref. no. **HUM106-G-FEDER**; and DÍGAME, ref. no. **JA.A1.3-06**). The research reported in this study has been funded by MCINAEI/ 10.13039/501100011033 (grant ref. no. **PRE2021-098899**).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

Data Availability Statement: The dataset used in this paper is the open access corpus Santa Barbara Corpus of American English. This corpus can be accessed and downloaded at the following url: https://www.linguistics.ucsb.edu/research/santa-barbara-corpus

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; Mcleavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In Proceedings of the Proceedings of the 40th International Conference on Machine Learning, PMLR; PMLR, 2023; pp. 28492–28518.
2. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Adv Neural Inf Process Syst* **2020**, *2020-December*.
3. Besdouri, F.Z.; Zribi, I.; Belguith, L.H. Arabic Automatic Speech Recognition: Challenges and Progress. *Speech Commun* **2024**, *163*, 103110, doi:10.1016/j.specom.2024.103110.
4. Hsu, M.-H.; Huang, K.P.; Lee, H. Meta-Whisper: Speech-Based Meta-ICL for ASR on Low-Resource Languages. **2024**.
5. Synnaeve, G.; Xu, Q.; Kahn, J.; Likhomanenko, T.; Grave, E.; Pratap, V.; Sriram, A.; Liptchinsky, V.; Collobert, R. End-to-End ASR: From Supervised to Semi-Supervised Learning with Modern Architectures. In Proceedings of the ICML 2020 Workshop on Self-supervision in Audio and Speech; 2020.
6. Aldarmaki, H.; Ullah, A.; Ram, S.; Zaki, N. Unsupervised Automatic Speech Recognition: A Review. *Speech Commun* 2022, *139*, 76–91.
7. Ramírez, J.; Górriz, J.M.; Segura, J.C. Voice Activity Detection. Fundamentals and Speech Recognition System Robustness. In *Robust Speech Recognition and Understanding*; Grimm, M., Kroschel, K., Eds.; InTech: Rijeka, Croatia, 2007; pp. 1–22.
8. Levinson, S.E.; Rabiner, L.R.; Sondhi, M.M. An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell System Technical Journal* **1983**, *62*, 1035–1074, doi:10.1002/j.1538-7305.1983.tb03114.x.
9. Allen, J.; Hunnicutt, S.; Carlson, R.; Granstrom, B. MITalk-79: The 1979 MIT Text-to-Speech System. *J Acoust Soc Am* **1979**, *65*, S130–S130, doi:10.1121/1.2017051.
10. Taylor, P.; Black, A.W. Assigning Phrase Breaks from Part-of-Speech Sequences. *Comput Speech Lang* **1998**, *12*, 99–117, doi:10.1006/csla.1998.0041.
11. Torres, H.M.; Gurlekian, J.A.; Mixdorff, H.; Pfitzinger, H. Linguistically Motivated Parameter Estimation Methods for a Superpositional Intonation Model. *EURASIP J Audio Speech Music Process* **2014**, *2014*, 1–13, doi:10.1186/s13636-014-0028-3.
12. Zen, H.; Senior, A.; Schuster, M. Statistical Parametric Speech Synthesis Using Deep Neural Networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; IEEE, May 2013; pp. 7962–7966.
13. Graves, A.; Mohamed, A.; Hinton, G. Speech Recognition with Deep Recurrent Neural Networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; IEEE, May 2013; pp. 6645–6649.
14. Roger, V.; Farinas, J.; Pinquier, J. Deep Neural Networks for Automatic Speech Processing: A Survey from Large Corpora to Limited Data. *EURASIP J Audio Speech Music Process* **2022**, *2022*, 19, doi:10.1186/s13636-022-00251-w.
15. Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.-Y. FastSpeech: Fast, Robust and Controllable Text to Speech. **2019**. Advances in Neural Information Processing Systems, 32, doi:10.48550/arXiv.1905.09263.
16. Juang, B.H.; Rabiner, L.R. Hidden Markov Models for Speech Recognition. *Technometrics* **1991**, *33*, 251–272, doi:10.1080/00401706.1991.10484833.
17. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Mohamed, A.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process Mag* **2012**, *29*, 82–97, doi:10.1109/MSP.2012.2205597.
18. Yu, D.; Deng, L. *Automatic Speech Recognition*; Springer London: London, 2015; ISBN 978-1-4471-5778-6.
19. Ning, H.; Liu, M.; Tang, H.; Huang, T. A Spectral Clustering Approach to Speaker Diarization. In Proceedings of the INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP; International Speech Communication Association, 2006; pp. 2178–2181.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

20. Piat, M.; Fohr, D.; Illina, I. Foreign Accent Identification Based on Prosodic Parameters. In Proceedings of the Interspeech 2008; ISCA: ISCA, September 22 2008; pp. 759–762.
21. Kocharov, D.; Kachkovskaia, T.; Skrelin, P. Prosodic Boundary Detection Using Syntactic and Acoustic Information. *Comput Speech Lang* **2019**, *53*, 231–241, doi:10.1016/j.csl.2018.07.001.
22. Hogg, A.O.T.; Evers, C.; Naylor, P.A. Speaker Change Detection Using Fundamental Frequency with Application to Multi-Talker Segmentation. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, May 2019; pp. 5826–5830.
23. Mehrish, A.; Majumder, N.; Bharadwaj, R.; Mihalcea, R.; Poria, S. A Review of Deep Learning Techniques for Speech Processing. *Information Fusion* **2023**, *99*, doi:10.1016/j.inffus.2023.101869.
24. Chorowski, J.; Bahdanau, D.; Cho, K.; Bengio, Y. End-to-End Continuous Speech Recognition Using Attention-Based Recurrent Nn: First Results. In Proceedings of the NIPS 2014 Workshop on Deep Learning; 2014.
25. Sell, G.; Garcia-Romero, D. Speaker Diarization with Plda I-Vector Scoring and Unsupervised Calibration. In Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT); IEEE, December 2014; pp. 413–417.
26. Wan, L.; Wang, Q.; Papir, A.; Moreno, I.L. Generalized End-to-End Loss for Speaker Verification. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, April 2018; pp. 4879–4883.
27. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, April 2018; pp. 5329–5333.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Advances in Neural Information Processing Systems* **2017**.
29. Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient Transformers: A Survey. *ACM Computing Surveys* **2023**, *55*, 1–28, doi:10.1145/3530811.
30. Roll, N.; Graham, C.; Todd, S. PSST! Prosodic Speech Segmentation with Transformers. **2023**, doi:10.48550/arXiv.2302.01984.
31. Taylor, P. *Text-to-Speech Synthesis*; 1st ed.; Cambridge University Press: New York, 2009.
32. Volín, J.; Řezáčková, M.; Matouřek, J. Human and Transformer-Based Prosodic Phrasing in Two Speech Genres. In Proceedings of the Speech and Computer. SPECOM 2021.; Karpov, A., Potapova, R., Eds.; Springer: Cham, 2021; pp. 761–772.
33. Lin, B.; Wang, L.; Feng, X.; Zhang, J. Joint Detection of Sentence Stress and Phrase Boundary for Prosody. In Proceedings of the Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH; International Speech Communication Association, 2020; Vol. 2020-October, pp. 4392–4396.
34. Hruz, M.; Zajic, Z. Convolutional Neural Network for Speaker Change Detection in Telephone Speaker Diarization System. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, March 2017; pp. 4945–4949.
35. Kwon, S.; Narayanan, S.S. Speaker Change Detection Using a New Weighted Distance Measure. In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002); ISCA: ISCA, September 16 2002; pp. 2537–2540.
36. Aronowitz, H.; Zhu, W. Context and Uncertainty Modeling for Online Speaker Change Detection. In Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, May 2020; pp. 8379–8383.
37. Snyder, D.; Garcia-Romero, D.; Sell, G.; McCree, A.; Povey, D.; Khudanpur, S. Speaker Recognition for Multi-Speaker Conversations Using X-Vectors. In Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, May 2019; pp. 5796–5800.
38. Fujita, Y.; Kanda, N.; Horiguchi, S.; Xue, Y.; Nagamatsu, K.; Watanabe, S. End-to-End Neural Speaker Diarization with Self-Attention. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU); IEEE, December 2019; pp. 296–303.
39. Anidjar, O.H.; Estève, Y.; Hajaj, C.; Dvir, A.; Lapidot, I. Speech and Multilingual Natural Language Framework for Speaker Change Detection and Diarization. *Expert Systems with Applications* **2023**, *213*, 119238, doi:10.1016/j.eswa.2022.119238.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 February 2025 doi:10.20944/preprints202502.0579.v1

40. Mateju, L.; Kynych, F.; Cerva, P.; Malek, J.; Zdansky, J. Overlapped Speech Detection in Broadcast Streams Using X-Vectors. In Proceedings of the Interspeech 2022; ISCA: ISCA, September 18 2022; pp. 4606–4610.
41. Bullock, L.; Bredin, H.; Garcia-Perera, L.P. Overlap-Aware Diarization: Resegmentation Using Neural End-to-End Overlapped Speech Detection. In Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, May 2020; pp. 7114–7118.
42. Du, Z.; Zhang, S.; Zheng, S.; Yan, Z. Speaker Overlap-Aware Neural Diarization for Multi-Party Meeting Analysis. **2022**. arXiv preprint arXiv:2211.10243.
43. Su, H.; Zhao, D.; Dang, L.; Li, M.; Wu, X.; Liu, X.; Meng, H. A Multitask Learning Framework for Speaker Change Detection with Content Information from Unsupervised Speech Decomposition. In Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, May 23 2022; pp. 8087–8091.
44. Berkling, K.; Zissman, M.A.; Vonwiller, J.; Cleirigh, C. Improving Accent Identification through Knowledge of English Syllable Structure. In Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 1998); ISCA: ISCA, November 30 1998; p. paper 0394-0.
45. Too Chen; Chao Huang; Chang, E.; Jingehan Wang Automatic Accent Identification Using Gaussian Mixture Models. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01.; IEEE, 2001; pp. 343–346.
46. O'Shaughnessy, D. Spoken Language Identification: An Overview of Past and Present Research Trends. *Speech Commun* **2025**, *167*, 103167, doi:10.1016/j.specom.2024.103167.
47. Watanabe, C.; Kameoka, H. GE2E-AC: Generalized End-to-End Loss Training for Accent Classification. **2024**. arXiv preprint arXiv:2407.14021.
48. Huang, H.; Xiang, X.; Yang, Y.; Ma, R.; Qian, Y. AISpeech-SJTU Accent Identification System for the Accented English Speech Recognition Challenge. In Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, June 6 2021; pp. 6254–6258.
49. Lesnichaia, M.; Mikhailava, V.; Bogach, N.; Lezhenin, I.; Blake, J.; Pyshkin, E. Classification of Accented English Using CNN Model Trained on Amplitude Mel-Spectrograms. In Proceedings of the Interspeech 2022; ISCA: ISCA, September 18 2022; pp. 3669–3673.
50. Matos, A.; Araújo, G.; Junior, A.C.; Ponti, M. Accent Classification Is Challenging but Pre-Training Helps: A Case Study with Novel Brazilian Portuguese Datasets. In Proceedings of the Proceedings of the 16th International Conference on Computational Processing of Portuguese; 2024; pp. 364–373.
51. Subhash, D.; G., J.L.; B., P.; Ravi, V. A Robust Accent Classification System Based on Variational Mode Decomposition. *Engineering Applications of Artificial Intelligence* **2025**, *139*, 109512, doi:10.1016/j.engappai.2024.109512.
52. Song, T.; Nguyen, L.T.H.; Ta, T.V. MPSA-DenseNet: A Novel Deep Learning Model for English Accent Classification. *Computer Speech & Language* **2025**, *89*, 101676, doi:10.1016/j.csl.2024.101676.
53. Viglino, T.; Motlicek, P.; Cernak, M. End-to-End Accented Speech Recognition. In Proceedings of the Interspeech 2019; ISCA: ISCA, September 15 2019; pp. 2140–2144.
54. Ghorbani, S.; Hansen, J.H.L. Advanced Accent/Dialect Identification and Accentedness Assessment with Multi-Embedding Models and Automatic Speech Recognition. *The Journal of the Acoustical Society of America* **2023**, 155(6), 3848-3860.
55. Ravanelli, M.; Zhong, J.; Pascual, S.; Swietojanski, P.; Monteiro, J.; Trmal, J.; Bengio, Y. Multi-Task Self-Supervised Learning for Robust Speech Recognition. In Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, May 2020; pp. 6989–6993.
56. Pascual, S.; Ravanelli, M.; Serrà, J.; Bonafonte, A.; Bengio, Y. Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks. **2019**. arXiv preprint arXiv:1904.03416.
57. Kunešová, M.; Zajíc, Z. Multitask Detection of Speaker Changes, Overlapping Speech and Voice Activity Using Wav2vec 2.0. In Proceedings of the ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, June 4 2023; pp. 1–5.
58. Zhang, J.; Peng, Y.; Van Tung, P.; Xu, H.; Huang, H.; Chng, E.S. E2E-Based Multi-Task Learning Approach to Joint Speech and Accent Recognition. **2021**. arXiv preprint arXiv:2106.08211.
59. Yolwas, N.; Meng, W. JSUM: A Multitask Learning Speech Recognition Model for Jointly Supervised and Unsupervised Learning. *Applied Sciences (Switzerland)* **2023**, *13*, doi:10.3390/app13095239.

60. Wang, R.; Sun, K. TIMIT Speaker Profiling: A Comparison of Multi-Task Learning and Single-Task Learning Approaches. **2024**. arXiv preprint arXiv:2404.12077.
61. Shah, S.M.; Moinuddin, M.; Khan, R.A. A Robust Approach for Speaker Identification Using Dialect Information. *Applied Computational Intelligence and Soft Computing* **2022**, *2022*, 1–16, doi:10.1155/2022/4980920.
62. Du Bois, J.W.; Chafe, W.L.; Meyer, C.; Thompson, S.A.; Martey, N. Santa Barbara Corpus of Spoken American English. *CD-ROM. Philadelphia: Linguistic Data Consortium* **2000**.
63. Kunešová, M.; Řezáčková, M. Detection of Prosodic Boundaries in Speech Using Wav2Vec 2.0. In Proceedings of the Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer Science and Business Media Deutschland GmbH, 2022; Vol. 13502 LNAI, pp. 377–388.
64. Labov, W.; Ash, S.; Boberg, C. *The Atlas of North American English*; Mouton de Gruyter, 2006; ISBN 978-3-11-016746-7.
65. Kashif, K.; Alwan, A.; Wu, Y.; De Nardis, L.; Di Benedetto, M.G. MKELM Based Multi-Classification Model for Foreign Accent Identification. *Heliyon* **2024**, *10*, doi:10.1016/j.heliyon.2024.e36460.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.